CLAIMS

That which is claimed:

1. A method for information extraction, comprising:

accessing a plurality of related articles;

determining a seed article from the related articles; and

identifying at least one information field within the seed article by comparing the seed article to at least one other related article.

2. The method of claim 1, further comprising:

determining a label for the information field; and

associating a pointer to a location of the information field in the seed article to create a template.

3. The method of claim 1, wherein comparing the seed article to at least one other related article is performed by a dynamic programming alignment algorithm to determine an alignment between the seed article and the related article.

4. The method of claim 1, further comprising determining a cluster of related articles from the related articles.

5. The method of claim 4, wherein determining a cluster of related articles is performed by using a dynamic programming alignment algorithm to compute edit

distances between the seed article and all of the related articles and choosing the cluster of articles based on the edit distances.

6. The method of claim 4, wherein the identifying at least one information field within the seed article is performed by comparing the seed article to the cluster of articles.

7. The method of claim 1, wherein the information field corresponds to variable data.

8. The method of claim 1, wherein the articles are web pages.

9. The method of claim 8, wherein the related articles are web pages on a web site.

10. The method of claim 9, further comprising simplifying the content on a web page.

11. The method of claim 10, wherein simplifying the content includes preserving visible text, visible images, and visible paragraph and table formatting.

12. The method of claim 2, further comprising:

identifying a plurality of templates each comprising at least one information field;

comparing a source article to the templates to determine the closest template;

associating data from the article with an information field from the closest template; and

extracting the associated data.

13. A method of extracting data from a source article, comprising:

identifying a plurality of templates each comprising at least one information field;

comparing the source article to the templates to determine the closest template; and

associating data from the article with an information field from the closest template.

14. The method of claim 13 further comprising extracting the associated data.

15. The method of claim 13, wherein comparing the source article to the templates is performed by a dynamic programming alignment algorithm to compute an edit distance between the source article and the templates.

16. The method of claim 13, wherein the source article is a web page.

17. A computer readable medium containing program code, comprising:

program code for receiving a plurality of related articles;

program code for determining a seed article from the related articles; and

program code for identifying at least one information field within the seed article by comparing the seed article to at least one other related article.

18. The computer readable medium of claim 17, further comprising:

program code for determining a label for the information field; and

program code for associating a pointer to a location of the information field in the seed article to create a template.

19. The computer readable medium of claim 17, wherein comparing the seed article to at least one other related article is performed by a dynamic programming alignment algorithm to determine an alignment between the seed article and the related article.

20. The computer readable medium of claim 17, further comprising program code for determining a cluster of related articles from the related articles.

21. The computer readable medium of claim 20, wherein determining the cluster of related articles is performed by using a dynamic programming alignment algorithm to compute edit distances between the seed article and all of the related articles and choosing the cluster of related articles based on the edit distances.

22. The computer readable medium of claim 20, wherein the identifying at least one information field within the seed article is performed by comparing the seed article to the cluster of articles.

23. The computer readable medium of claim 18, further comprising:

program code for identifying a plurality of templates each comprising at least one information field;

program code for comparing a source article to the templates to determine the closest template;

program code for associating data from the article with an information field from the closest template; and

program code for extracting the associated data.

24. The computer readable medium of claim 23, wherein comparing the source article to the templates is performed by a dynamic programming alignment algorithm to compute an edit distance between the source article and the templates.